

Regrouper ses données de recherche dans une base de données relationnelle

Une brève introduction pour des collectes informelles

Les fondamentaux

- Une **donnée** est une représentation d'un aspect de la réalité. Elle est partielle et située, mais, quand elle est contextualisée, elle fournit une **information** qui peut être exploitée pour construire de la **connaissance**.
- Une base de données est un regroupement de données. Une **base de données relationnelle** établit des relations entre les données qu'elle contient.
- Une base de données s'appuie sur des **tables**. Chaque table correspond à un aspect de la réalité que les données qu'elle réunit représentent.

Bonnes pratiques

Pour limiter la perte de données et permettre leur exploitation, il faut que la base de données soit rigoureusement organisée, notamment en s'assurant de la cohérence de son agencement et de l'harmonie des valeurs qu'elle contient.

- Assurer le maintien de l'**intégrité référentielle** : chaque référent doit avoir un et un seul signifiant.

Pour la machine, "Gabrielle Roy" et "Roy Gabrielle" sont deux références différentes. Cela vaut pour les valeurs, mais aussi pour les noms de variables : "Auteur" désigne la même variable dans la table oeuvres et dans la table auteurs.

- Ajouter une ou des **tables intermédiaires** comportant une ligne par relation (entre des éléments de tables différentes) avec en valeurs celles des tables d'origines, plutôt que de lister les différentes correspondances sur une même ligne.

Une oeuvre peut avoir plusieurs auteurices et une auteurice peut avoir écrit plusieurs oeuvres. Dans ce cas, il faut ajouter aux tables oeuvres (fournissant des informations sur les oeuvres) et auteurs (fournissant des informations sur les auteurices) une table oeuvres_auteurs listant l'ensemble des relations existant entre chaque oeuvre et chaque auteurice.

- **Documenter** les choix effectués : tenir un fichier .txt avec les différents problèmes qui apparaissent lors de l'alimentation de la base et les réponses apportées. Préciser quelles sont les tables, leurs variables et les valeurs que celles-ci acceptent ou non.

La variable "nom_auteur" correspond-elle au nom de plume ? Faut-il distinguer Romain Gary et Émile Ajar ? Faut-il mettre le prénom ou le nom en premier ? etc.

Éviter

oeuvres

Titre	Auteur·ice·s	Date
Au Château d'Argol	Louis Poirier	1938
Graal théâtre	Delay Florence, Roubaud Jacques	1977-2005
Candide	Voltaire	1759

Préférer

Readme

Cette base de données regroupe les oeuvres sur lesquelles je travaille dans le cadre de ma thèse. Elle est constituée de la table **oeuvres** qui répertorie les oeuvres, de la table **auteurs** qui répertorie leurs auteurs et de la table intermédiaire **oeuvres_auteurs**.

"Titre" correspond au titre de l'oeuvre, le plus complet possible.

"Date" correspond à l'année (YYYY) de première publication de l'oeuvre dans son ensemble (paratexte exclu).

"Auteur" correspond au nom de plume de l'auteur, sous la forme Prénom Nom.

"Nom" correspond au nom de l'état civil de l'auteur au moment de sa mort, ou en 2024.

"Prénom" correspond au prénom principal de l'auteur selon son état civil, au moment de sa mort ou en 2024.

Tous les caractères spéciaux sont conservés.

"ID_oeuvres_auteurs" est un identifiant unique correspondant à une chaîne de caractères représentant un nombre. Le nombre en question est insignifiant et sert uniquement à désambiguer le référent.

Cette base de données a été créée le 2024-01-01 et a été mise à jour le 2024-11-19.

oeuvres_auteurs

ID_oeuvres_auteurs	Titre	Auteur
1	Au Château d'Argol	Julien Gracq
2	Graal théâtre	Florence Delay
3	Candide ou l'Optimisme	Voltaire
4	Graal théâtre	Jacques Roubaud

oeuvres

Titre	Date
Au Château d'Argol	1938
Graal théâtre	2005
Candide ou l'Optimisme	1759

auteurs

Auteur	Nom	Prénom
Julien Gracq	Poirier	Louis
Florence Delay	Delay	Florence
Voltaire	Arouet	François-Marie
Jacques Roubaud	Roubaud	Jacques